Corpus Linguistics for the Annotation Manager

Karen Fort, Adeline Nazarenko, Claire Ris

Corpus Linguistics for the Annotation Manager

Karën Fort[*] [**], *Adeline Nazarenko*[*], *Claire Ris*[**]
[*] LIPN – Paris 13 University & CNRS
Villetaneuse, FRANCE
[**] INIST – CNRS
Vandoeuvre-lès-nancy, FRANCE
adeline.nazarenko@lipn.univ-paris13.fr
{karen.fort,claire.ris}@inist.fr

Abstract

Hand crafted annotated corpora are acknowledged as critical elements for the Human Language Technologies but systems have to be trained on domain specific data to achieve a high level of performance. This is the reason why numerous annotation campaigns are launched. The role of the annotation manager consists in designing the annotation protocol, sometimes selecting the source data, hiring the required number of annotators with the adequate competences, writing the annotation guidelines, controlling the annotation process and delivering the resulting annotated corpus with the expected quality.

However, for a given task, the complexity of the annotation work seems to be highly dependent on the type of corpus to annotate. Since this affects both the cost and the quality of the annotation, it is an important issue to tackle for the annotation manager.

This paper illustrates the role of corpus linguistics for the management of annotations through a specific annotation campaign. We show how the corpus characteristics affect all aspects of the annotation protocol: the design of the annotation guidelines, the selection of the a sub-corpus for training, the duration of the annotator's training, the complexity of the annotation formalism, the quality of the resulting annotation.

1 Introduction

The term "annotation" corresponds to the process of adding (*ad-*) an interpretation in the form of a *note* on a flow of data. This definition stems from and enlarges the one given in (Leech, 1997) as the interpretative dimension of the annotation is not limited to linguistic information or to any particular type of corpus, which can be composed of speech and texts, but also of video or images.

Since the beginning of the 90s and the seminal work on the Penn Treebank project (Marcus *et al.*, 1993), there has been a large endeavor to develop various types of annotated corpora, which are used for collecting linguistic data and testing linguistic theories but also and probably now more often, for training, testing and evaluating Natural Language Processing (NLP) tools.

Such annotated corpora are usually developed through annotation tasks in which one or several human annotators are asked to encode their interpretation of a given corpus in the form of annotations that are attached to that corpus. However, such annotation tasks are complex to define. What kind of interpretation is required? for what purpose? How many annotations are needed? What type of corpus and which corpora have to be annotated? What quality of annotation is expected? Those are the typical questions that should be answered when defining an annotation task in order to 1) write down annotation guidelines that explain to the annotators what kind of interpretation they are expected to deliver, 2) choose an annotation tool, 3) determine how many annotators should work in parallel, and 4)the size of the corpus to be annotated.

Even if, to date, there is no stable, well-acknowledged methodology for designing such an annotation task, it is more and more evident that the annotation manager plays a central role in this process. The growing need for annotated corpora has given rise to a new job.

The *annotation manager* has a critical role in the design and success of the annotation tasks, for which s/he is responsible. S/He has to interact with the various actors involved in the annotation process, mainly the *client*, who asks for an annotated corpus, and the *annotators*, who annotate the corpus. S/He has to understand the client's needs, estimate the costs, write the annotation guidelines, hire and train the annotators and evaluate the quality of the annotations before they are delivered to the client.

Corpus linguistics is important for the management of annotations, not only for the selection of the corpus to annotate – quite often the sources to annotate are selected by the client and the annotation manager has no control on the corpus choice itself – but because the corpus imposes strong constraints on the annotation task, which cannot be defined independently of the availability, size, homogeneity and internal characteristics of the corpus. The annotation manager cannot take the responsibility of an annotation task without an in-depth analysis of the corpus to be annotated.

This paper illustrates the role of corpus linguistics for the management of annotations through a specific annotation campaign. The task itself is introduced in Section 2 and the underlying corpus analysis is presented in Section 3. Section 4 explains what is the role of the annotation manager in this context and how much it relies on corpus analysis.

2 Football Matches Annotation Task

The annotation of football (soccer) matches reports is a task that was defined in the context of the Quæro program[1] in which various Quæro partners are involved (IRISA, LIPN, INIST, among others):

– The client who expressed the need for an annotated corpus is both the primary user and the provider of the corpus.[2] The source documents – the corpus – was provided by the same client with the summarization of matches reports and various other re-purposing tasks as target applications.

– The annotations were made at INIST-CNRS. The campaign was organized there by an annotation manager (Karën Fort) with few expert annotators (3 at the beginning and then 2), using an annotation interface, Glozz (Widlocher *et al.*, 2009).

– The *a priori* corpus analysis was made jointly by INIST and LIPN, and the *a posteriori* error analysis by INIST, LIPN and IRISA (Vincent Claveau).

Figure 1 shows the role of the annotated corpus in the summarization application. The goal is to produce summaries out of the football match reports that are given as input. In the training phase (bottom part of the figure), the summarization tool is given couples of reports and associated summaries and "learns" how to derive the latter from the former. In the exploitation phase (upper part of the figure), the resulting derivation rules are exploited to derive new summaries for unknown match reports.

3 Corpus Analysis

In this task, the annotation manager had no control on the selection and composition of the source corpus that were made by the client. However, it is essential to take corpus characteristics into account in the definition of the annotation task.

1. `http://quaero.org/`
2. IRISA, team TexMex (Vincent Claveau).

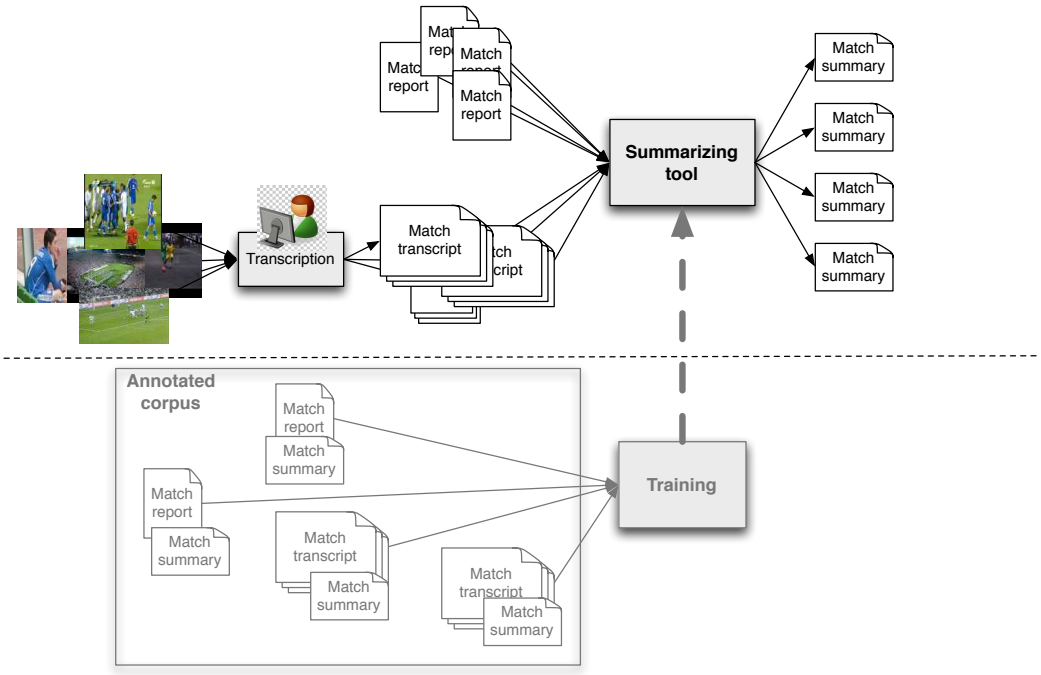

Figure 1. Role of the annotations for the training of a summarization tool

First important characteristic, two types of match reports are taken into consideration: written minutes produced through a web application and transcripts of match video commentaries. The composition of the corpus is the following:

– 12 football matches in French

- 12 written minutes (Web)
- 24 transcripts of the video commentaries (1 file per half-time)

– Written minutes of 4 additional matches

Despite the multimedia dimension of the source corpus, the annotation is defined as text-based. No access to source videos was given to the annotators since the final summarization tool would have to rely on texts only.

The source corpus also includes various types of matches as shown on Table 1. This diversity has a direct impact on names and denominations, which differ at the national and international level. For instance, out of context, the mention *les joueurs de l'équipe de France / the French team players* is difficult to interpret. One does not know if it

refers to the French national team as such (at the international level) or to the French team players that are playing for a specific club. For instance, Anelka is playing both in the French Team and for Chelsea.

Match types	National Level		International Level	
	French national level (Ligue 1)	Other national level (Bundesliga)	Club teams (Champions' League)	National teams (World Cup)
Proportion	25%	8%	50%	17%
Example	Bordeaux-Vannes	WerderBreme -Hambourg	Munich-Lyon	France-Serbie

Table 1. Composition of the football corpus wrt. match types

A third important characteristic to take into account is the heterogeneity of the report size in the source corpus. This is reflected in the Table 2 and illustrated on Figure 2, which presents the same action in two different report types. In whole, the corpus is unbalanced: the minutes and transcripts respectively represent 16 % and 84 % of the total 247,955 tokens.

Types of report	Minimal length	Maximal length	Total	Proportion
Minutes	1,116 tokens	3,627 tokens	38,919 tokens	16%
Transcripts	6,020 tokens	11,110 tokens	209,036 tokens	84%

Table 2. Contrast between minutes and transcripts, the two types of football reports

A forth point to note is the variety of the commentary sources. Videos come from 5 different channels and are commented by 4 different teams. The minutes are typed by individual commentators in reverse chronological order (as in a blog). This heterogeneity is reflected in the quality and style of the resulting texts:

– There is a major contrast between the oral and written styles. The transcripts are verbose and include linguistic features that are typical of the oral such as hesitations, disfluencies, repetitions and errors (*e.g. Oh là là. le contrôle. le contrôle, le contrôle. / Oops. control. control ...* or *Bosingwa , Obi Mikel ... Terry ... Malouda ...*).

– There is also a contrast between the commentaries that are dialogues (*de toutes les manières Jean-Michel ... / anyway Jean-Michel...*) and the minutes that are monologues.

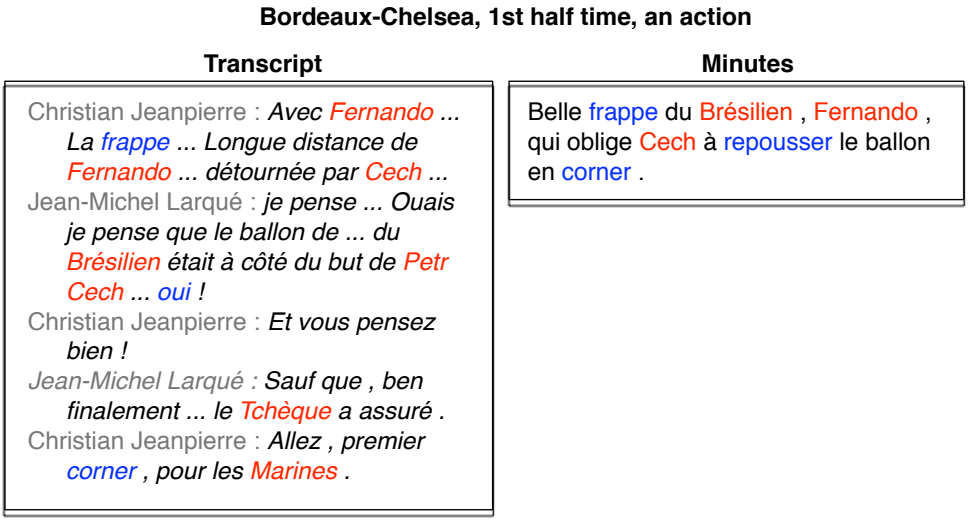

Figure 2. Comparison of the size of minutes and transcripts.

– Finally, there are noticeable differences among commentators' styles, each commentator having his preferred denominations (*Bordeaux, les girondins, les bordelais, les marines*, all referring to the same Bordeaux team players) and some being more explicit than others (especially in the description of actions).

4 The Role of the Annotation Manager

After analyzing the corpus to annotate, the annotation manager has to set up the annotation campaign. S/He must precisely define the task and design annotation guidelines, analyze the task complexity and estimate its costs, select and train the annotators and finally evaluate and control the quality of the produced annotations. The example of the football matches campaign shows that several of these tasks directly rely on the initial corpus analysis.

4.1 *Task and Guidelines Definition*

In the context of the football matches campaign, the aim is to exploit the annotations for various re-purposing applications such as the production of match summaries, the alignment of commentaries on videos, a statistical analysis of the players' activity. In

this context, the relevant information is the sequence of football actions and resulting scores (who does what at what time and for which result?) and the goal is to annotate it. The set of annotations made on a match report must form a summary of the match.

Finally, the annotation manager decided to ask for the following annotations:

Actors player, team, referee, assistant referee, coach, club president

Locations match location (town), field area

Chronology time in match

Actions center, center attempt, direct free kick, indirect free kick, corner, penalty, dribble, foul, offside, score goal, miss goal, stop goal, interception, have ball, yellow card, red card, warning, audience action

Relations pass, pass attempt, combination, tackle foul, replacement, foul

This task definition represents a trade-off between the richness of the annotation and the difficulty of the annotators' work. From the task definition, the annotation manager has to design the guidelines. Their role is to explain the task to the annotators and guide their work (what pieces of texts should be annotated and how?) but also to reduce the annotation errors and to standardize the annotations made by various annotators and over time. In the context of the football match campaign, it was decided to have a single guide and annotation task, despite the corpus heterogeneity. Since the source of reports is known and the training must be done separately for each type of sources, it would have been possible to carry out two different annotation campaigns. However, that would have been a significant additional burden for the annotators and a possible source of errors, if the same annotators had to tackle with two different but somehow similar annotation tasks. The manager therefore preferred the robustness of the guidelines to the accuracy of the annotations.

These decisions had a direct impact on the annotation model. It led to simplify it, so that annotators could annotate in the same way all the documents, while listing the specific cases. The annotation manager also decided to allow for commentaries in case the annotators were uncertain what or how to annotate.

The requirement of robustness imposed that football actions be annotated in the same way in video commentaries and in written minutes. However, video commentaries con-

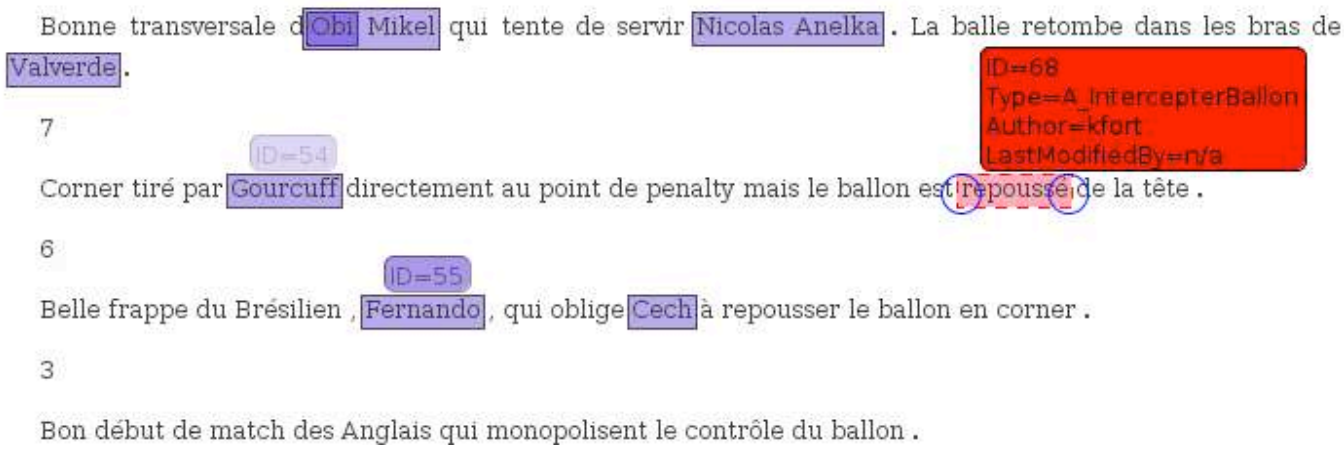

Figure 3. Example of action annotation with a missing actor

tain frequent ellipses. The annotation manager asked the annotators to annotate the actors of an action or a relation, not the action or relation itself, as action and relation markers are often omitted. For instance, in *Bosingwa. Le ballon va sortir. / Bosingwa. The ball goes out.*, The `shot` action is attached to the actor (*Bosingwa*). Another example is presented in Figure 3. If the actors themselves are omitted, the annotators are asked to annotate the action or relation marker and add a feature "missing actor" (source or/and target actor for relations). In *La balle est repoussée de la tête / The ball is pushed away with the head* , the `intercept` action is attached to the verb (*pushed*).

4.2 *Cost Estimates*

The annotation manager has to analyze the annotation task so as to precisely estimate its cost, potentially redefine it to reduce it and provide the annotators with the appropriate assisting tools, such as an automatic pre-annotation or a convenient annotation tool.

To do so, we propose to use five complexity indicators (Fort *et al.*, 2011), impacting the final cost of the task:

1) The *unit discrimination* is defined as the ratio of units to annotate with regard to the total number of “annotatable“ units, sparse annotation being difficult to produce.

2) The *frontiers adjustment* is defined with regard to a standard segmentation of the text and is related to the number of units which frontiers must be changed during the

annotation process.

3) The *annotation language expressiveness* is related to the type of annotation language that is used (the tags can be types, relations or even 2nd order annotations).

4) The *tagset dimension* captures the fact that the annotation work is more difficult when the degree of liberty in tagging and the number of candidate tags is higher.

5) The *degree of ambiguity* gives an estimate of the number of ambiguous units that can be tagged in different ways.

The complexity of the football campaign annotation we present is mainly related to unit discrimination, annotation language expressiveness and ambiguity.

4.2.1 *Unit Discrimination*

Independently of the number of annotations to produce, sparse annotations are more difficult than others. Annotation is easier if all the units have to be annotated or if the relevant ones were pre-tagged, but this is not the case in the football task which has, on the contrary, a high level of discrimination. Since the annotation is a summary, only a small proportion of transcripts has to be annotated and this discrimination complexity is higher for video transcripts, which are verbose.

To ease the work of the annotators, the annotation manager decided to pre-annotate the corpus. Some actors (proper names of players and coaches) were thus automatically pre-annotated. Unfortunately, not all the actors could be pre-annotated: the names of referees and club presidents, as well as variant denominations remained unannotated. For instance, *Fernando* is pretagged but not *the brazilian*.

4.2.2 *Annotation Language Expressiveness*

The annotation task complexity depends on the type of information that the annotators have to add. The annotation is boolean if the units are simply marked as relevant or not. Type annotation refers to the cases where a simple tag is associated to units. Part of the annotations of the football campaign belong to that category (player/coach/referee/match location) even if multi-type annotations are also expected in the case, for instance, where a name refers both to a player and to the captain. The complexity is higher for relational annotations (*X makes a foul on Y*) and for meta annotation (annotation on an existing annotation).

To reduce the expressiveness of the annotation language in the football campaign, the annotation manager simplified the annotation task. Actions and relations were all annotated as types on the actors. Although this might appear counterintuitive at first, it eases the annotators' work. Once the annotators understood a football action, they were able to annotate its main actor, who is usually easy to identify (if not pre-annotated). The annotators do not have to localize the exact words that express the action or the target of the relation, which would be difficult, due to the elliptic style of the reports.

The annotation manager also asked the annotators to indicate the most dubious annotations, an important information for controlling the quality of the annotations (see Section 4.3), especially in complex and non-standard annotation campaigns as the football one. However, to keep the annotation language as simple as possible, it was decided that the uncertainty would be encoded as an annotation feature rather than as a meta-annotation apposed on dubious annotations.

4.2.3 *Degree of Ambiguity*

The degree of ambiguity is the complexity indicator that is the most difficult to compute. One has to estimate how many alternative tags a given text unit may have. Even if it is not possible to produce precise figures, it is easy to understand that a high degree of ambiguity characterizes the football task. Since actors, actions and relations annotations are all attached to actors, it increases their ambiguity. In addition to that, many player and team denominations are ambiguous and ellipses even increase ambiguity.

The degree of ambiguity remains the most important factor of complexity in the football task. As a matter of fact, the choices made by the annotation manager in designing the guidelines aimed at reducing the discrimination and the expressiveness complexity, but led to increasing ambiguity, which annotators had to tackle.

4.3 ***Evaluation***

The last role of the annotation manager is to control the quality of the annotations.

4.3.1 *Protocol*

The annotation manager organized the campaign in such a way that the annotators' training and learning curve, both fundamental to the annotation quality (Marcus *et al.*, 1993, Dandapat *et al.*, 2009), were taken into account.

The campaign encompassed three different phases. In the *training phase*, the annotators annotated a small set of documents. They got used to the corpus, task, guidelines and annotation tool. The annotation manager revised the guidelines taking into account the annotators' feedback, monitored the annotation quality and speed, which were increasing. In the football campaign, the training phase also led to eliminating one of the annotators (he/she was not reliable enough).

In the *pre-campaign*, the annotation manager dispatched the work to the annotators, who annotated the same sample of corpus, in order to compute the inter-annotator agreement. The annotation manager performed a detailed analysis of the disagreements, uncertainty features and annotators' feedback and revised the guidelines accordingly. In parallel, the annotation manager monitored the annotation speed, which was still increasing. Once it stabilized, the annotation manager put an end to the pre-campaign phase.

In the production phase, the annotation manager dispatched the work to annotators, who worked independently of each other. The annotation manager controlled the quality of the annotation.

4.3.2 *Quality Evaluation*

Following the methodology described in (BM *et al.*, 2005), the inter-annotator agreement was computed early in the process (in the training phase) in order to assess the reliability of the annotations. The guidelines were then revised to reduce the disagreement sources. The intra-annotator agreement was also computed during the campaign to check the annotators' ability to reproduce the annotations (Krippendorff, 2004).

The coefficients used to compute the agreements are usually those of the kappa family (Cohen's or Carletta's, see (Artstein *et al.*, 2008) for more details), which take chance into account, as opposed to simpler measures like the F-measure. However, the computation of those coefficients requires to know the number of "annotatable" units, that can only be estimated for tasks like this (Grouin *et al.*, 2011). Other measures are also being considered, like the one provided in the annotation tool (Mathet *et al.*, 2011).

A detailed analysis of the obtained results should then be done, in order to identify the patterns of disagreements and correct the annotations and the guidelines. The annotators' feedback, given through uncertainty features or comments, should also be analyzed and taken into account.

5 Conclusion

The growing need for annotated corpora gave rise to a new job, the annotation manager and it is now a challenging task to set up a stable, well-acknowledged methodology to design such annotation tasks and, more generally, for the management of annotations.

This paper reports on an experiment where a heterogeneous corpus of football match reports had to be annotated in order to identify the key actions of the football players during the matches. This showed that corpus linguistics is important for the management of annotations, not only for the selection of the corpus to annotate – quite often the sources to annotate are selected by the client and the annotation manager has no control on the corpus choice itself – but because the corpus imposes strong constraints on the annotation task, which cannot be defined independently of the availability, size, homogeneity and internal characteristics of the corpus. The annotation manager cannot take the responsibility of an annotation task without an in-depth analysis of the corpus to be annotated.

6 Acknowledgments

We want to thank Alain Zérouki from INIST-CNRS, for his hard work as annotator and Vincent Claveau, for hir implication and help during and after the campaign. This work was realized as part of the Quæro Programme[3], funded by OSEO, French State agency for innovation.

3. http://quaero.org/

7 References

Artstein R. and M. Poesio. (2008) 'Inter-Coder Agreement for Computational Linguistics', *Computational Linguistics*, vol. 34, num. 4, 555–596, MIT Press.

Bonneau-Maynard H., S. Rosset, C. Ayache, A. Kuhn and D. Mostefa (September (2005)) Semantic Annotation of the French Media Dialog Corpus. *InterSpeech*, *Lisboa, Portugal*, *2005*.

Dandapat S., P. Biswas, M. Choudhury and K. Bali (2009) Complex Linguistic Annotation - No Easy Way Out! A Case from Bangla and Hindi POS Labeling Tasks. *Proceedings of the third ACL Linguistic Annotation Workshop*, *Singapore*, *2009*.

Fort K., A. Nazarenko and S. Rosset. (2011) 'Manual Annotation of Corpora: Identifying the Difficulties to Reduce them', *Submitted*.

Grouin C., S. Rosset, P. Zweigenbaum, K. Fort, O. Galibert and L. Quintard (June (2011)) Proposal for an Extension of Traditional Named Entities: From Guidelines to Evaluation, an Overview. *Proceedings of the 5th Linguistic Annotation Workshop*, *Portland, Oregon, USA*, *2011*. Association for Computational Linguistics, pp. 92–100.

Krippendorff K. (2004). 'Content Analysis: An Introduction to Its Methodology, second edition', chapter 11. Thousand Oaks, CA., Sage.

Leech G. (1997). 'Corpus annotation: Linguistic information from computer text corpora', chapter Introducing corpus annotation, pp. 1–18. London, Longman.

Marcus M., B. Santorini and M. A. Marcinkiewicz. (1993) 'Building a large annotated corpus of English : The Penn Treebank', *Computational Linguistics*, vol. 19(2), 313-330.

Mathet Y. and A. Widlöcher (27 juin - 1er juillet (2011)) Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. *Actes de Traitement Automatique des Langues Naturelles 2011 (TALN 2011)*, *Montpellier, France*, *2011*.

Widlöcher A. and Y. Mathet (2009) La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. *Actes de Traitement Automatique des Langues 2009 (TALN 2009)*, *Senlis, France*, *2009*.